BERT Takes a BoW: Exploring natural language processing models for clinical diagnosis of neurodegenerative diseases

Aaisha Ameen

20854135

April 24, 2024

For the completion of BIOL 487 *Final Project*

Contents

Abstract

This work aims to explore the realm of natural language processing (NLP) models as a basis for supporting clinical diagnoses of neuropathological disorders (ND) that go on to be medically investigated, tracked and treated. The diverse clinical heterogeneity characteristic of NDs present challenges, both for humans and machines, to analyze patient signs and symptoms from a holistic, generalized yet specific perspective which creates a need for critical attention in developing relevant and effective NLP models. Based on a published recurrent neural network model that can predict the neuropathological disorder given common diagnoses, this work aims to explore two NLP models (bag-of-words, BoW, and Bidirectional-Encoder Representations from Transformers, BERT) from the perspective of comparing performances at a high level between two models. By changing the model used to classify the medical history data, it was determined that transformer-based models train better and can potentially perform better according to micro and sample f1-scores. The impact on training performance of BERT with deviations in completeness in dataset were considered from a model optimization standpoint as related to the incomplete nature of medical data from a clinical perspective. It was observed that greater number of randomized samples, the lower the observed training loss and the sooner the model converged to its minimal value. Considerations of activation functions are also discussed from the perspective of improving transformer models. Development of more brain-like language processing models is of great significance for ongoing research neurological disorders, and this work is a high-level exploration of this field.

Introduction

Clinical diagnoses often harbor ambiguous heterogeneity in their ability to describe signs and symptoms of neurodegenerative disorders clearly and divisively.

Artificial intelligence earned its moniker from the neurally-inspired roots from which it sprang forth; understanding how best to employ it to conduct tasks, process, generate, and "think" in applications that the human mind engages in is a longstanding journey. The motivation for understanding how to classify neuropathological diagnoses/attributes to clinical diagnoses is to support advances in medical sciences and ultimately health care when dealing with the complex situations of neuropathological patients.

By studying the outcomes of various neurodegenerative disorders, termed 'clinical disease trajectories' from the Netherlands Brain Bank (NBB) donors (n = 3042), Mekkes et al. develop a natural language processing (NLP) model capable of identifying 84 attributes for the classification of neurodegenerative diseases, as a resource and tool for investigating symptomatology, or the characterization of symptoms of a disorder exhibited by a patient. This model highlights their success by their ability to identify attributes differing between misdiagnosed disorders, which emphasizes a different kind of understanding and testability that machine learning methods can be used to supplement the development of medical understanding and discovery.

Transfer learning allows the reuse and repurposing of a pre-trained model to learn and solve new tasks, and 3 of the 5 employed transformer-based models in the Mekkes et al. study follow this approach. Medical data acquired at the Netherlands Brain Bank included neuropathological examinations, diagnosis histories, health statuses and familial conditions to derive labelled data (neurologically diseased, clinically diseased, non-disordered neuropathological traits, non-singly-classifiable clinical disorders, and psychiatric diagnoses (PD)) for training and testing their machine learning models. 18917 sentences worth of medical history data was scored, and 90 attributes were identified so that the clinical diagnoses could be matched to the Human Disease Ontologies. The team went on to study 5 different NLP methods to compare how accurate different frameworks are at achieving accuracy when performing across different machine learning scoring metrics.

Prior works (Alsetnzer et al. 2019; Lehman et al. 2023) have also reported the need for and importance of highly-specialized domains like clinical text based on their independent investigation of different NLP models performing on different clinical tasks. Alsentzer et al. surmise that there is a need for clinically-oriented BERT models to enhance performance of NLPs using clinical embeddings as they believe in the utility of domain-specific contextual embeddings for NLP tasks. Schrimpf et al. (2021) have also used AI to shed light on the predictive processing power of the brain – their exploration of different NLPs (including GPT-3) showed that the greater

the predictive capacity of the model, the more brain-like it was. They saw this by studying nodes in the networks and comparing the activities to fMRI data and intracranial ECG measurements in epileptic patients and observed shared patterns with respect to the speed at which text was read.

With the ultimate goal of building a multilabel classifier, training and employing NLP models was the focus of this study in order to see whether machines can understand language like humans. Motivated by the perspectives of clinicians, physicians or medical researchers, models such as the ones developed by Mekkes investigate whether they can count on the conclusions a learned machine makes by the way it finds patterns in data and generates summaries or infers information. Mekkes et al. wanted to see whether a computational pipeline relying on NLP models could standardize medical record summaries into clinical disease trajectories for 11 neuropathological disorders (described in Table 5) based on 90 signs and symptoms (which they term attributes in the scope of their models and are listed in Table 4) given the heterogeneity of clinical aspects of neuropathological disorders.

Multilabel classifiers are distinct from other classification problems in that they ascribe input samples to multiple categories (Devlin et al. 2019). When employed in text classification scenarios, they can be especially useful to classify inputs that don't fit exclusively in single categories, which is often a trait of signs and symptoms observed in patients enduring various neurological diseases. This model was complex and incredibly thorough for its very specific task of predicting clinical trajectories – for the scope of this course I decided to explore more simplified aspects of that aim from the perspective of language learning models using both clustering and machine learning methods.

In this preliminary look at NLP models used for clinical applications, we will delve into the details of the published models and test some basic parameters between two select models, bag-of-words (BOW) and a transformer-based BERT (Bidirectional-Encoder Representations from Transformers) variant (Gu et al. 2020). The scope of this study will explore the two following NLP frameworks used by Mekkes et al. while seeking to implement aspects of biological plausibility into chosen methods with the aim of understanding to what extent there is transferability between machines and humans learning how to naturally process language.

Bag-of-Words Modelling

This vectorization method relies on the following general workflow:

1. Data preprocessing:
    a. WordPiece tokenization (breaking up text into representative symbols)
    b. Word Lemmatization to identify base or roots of words
2. Splitting corpus into training and test sets

    a. Training employed in model fitting and evaluation (model validation) through by predicting training set outcomes
3. Encoding: involves labelling target variables to transform string data (which is a categorical data type) into numerical data (ordinal data type) which can be understood by the model
4. Sometimes: word vectorization in the event of feature extraction (such as in the case of the bag of words method)
5. Model training using ML algorithm to predict outcomes

BERT

BERT is an encoder-based feed-forward model utilizing only the encoder portion of a transformer and is particularly specialized at classification of sentences which is the basis of our model of focus. By incorporating the use of an attention layer, BERT is able to place words into context which is a limitation of BoW (Devlin et al. 2019).

BERT is pretrained on unsupervised tasks like masked language modelling (MLM) and next sentence prediction. During its pretraining stage, inputs are randomly masked (or omitted) with either special masking tokens, random tokens or the same token and the model is tasked with predicting the original sentence (Zanella & Toussaint 2022). Masking beginning or ending words is termed as padding and plays a role in homogenizing input tensor lengths. BERT models go a step further by predicting whether two sentences would follow consecutively after the other, highlighting this model's ability to grasp relevance based on natural language understanding (NLU).

BERT is built on multiple transformer encoder layers which use self-attention mechanisms to generate a richer picture of its input. Being bidirectional, this model learns the context of both preceding and succeeding information of each input word, which it first tokenizes using the WordPiece method, something that would be most appropriate for learning the semantics of out-of-vocabulary scientific nomenclature. When tokens have semantic context, they are called embeddings. These embeddings provide information about the token's position in the sentence, and this information is processed by the attention-headed Transformer-encoder layer. Tokenizers transform text into mathematical representations for the model's understanding, and whether this method of learning translates to biological learning or not is what groups like Goldstein, Schrimpf and Alsentzer et al. form the basis of their work. This pre-trained model is then fine-tuned to the task of focus (in this case text multilabel classification) leading to the tuning of weights from its pre-trained state.

BiomedBERT is a BERT variant specialized on PubMed and PubMed Central abstracts and articles; with its background training on peer-reviewed literature in the medical and life sciences,

it is easily adaptable to processing tasks that are specific to this domain and differentiating relevant concepts (Gu et al. 2020). Table 2 covers the specific configuration details highlighting the largeness and complexity of this deep learning model based on the size of its hidden layers.

Methods

Overview of Publication

The published study decided to employ 2 conventional NLP frameworks (BOW, SVC) and 3 pre-trained transformer-based NLP models (which they fine-tuned using their corpus of medical data after importing from HuggingFace). Performances were evaluated using the Scikit-learn classification report module.

After first parsing medical record summaries (via Python based parsers and the FuzzyWuzzy library) Mekkes et al. identified the attribute distribution for main diagnoses before labelling them to formal clinical diagnoses. They include a description of how accurate labels are for each attribute-disorder designation (on a scale from accurate to ambiguous).

Mekkes et al. use a multiple stratification method coupled with k-fold cross validation (k=5) of training for refinement on 80% of the data and validated through testing on the remaining 20% (Pedregosa et al. 2011). This supervised method ensured partitioning of the dataset in an equivalent attribute-distributive way so that training wasn't biased on one attribute over another. In summary, 90 attributes (Table 4) were identified out of roughly 200,000 parsed sentences before employing a stratified sampling strategy and k-fold cross validation (Trent 2022). This data was then employed in the following models described in Table 1.

k-fold cross validation is a resampling technique used to split input data into training and testing sets which is used to evaluate the robustness of a model's performance by splitting the dataset into k-number of groups, training on k-1 groups and testing on the remaining group (or fold). Each k^{th} fold is used once as the test set, so this process repeats k times, and the resultant performance metrics are averaged (f1-score, accuracies, etc.). This sampling technique is thought to reduce variability in the datasets as well as detect whether overfitting could be taking place (Sechidis et al. 2011).

Table 1. Text classification methods employed by Mekkes et al. (2024)

NLP Architecture/Algorithm	Description
Bag-of-words (BOW)	Baseline classification model
Support vector classification or machine (SVC, SVM)	Baseline classification
PubMedBERT	A transformer-based model pre-trained on PubMed abstracts

BioClinical_BERT	A transformer-based model pre-trained on electronic health records (EHRs)
T5	Text-to-text transformer using an encoder-decoder process to focus on transfer learning methods

This study compares the usage of a non-transformer-based bag-of-words method against the transformer-based BERT method, as well as some attempts at tuning different aspects of training**.** This work won't explore the problem-space to the full extent that the publication covers but will instead focus on comparing different aspects of two models when dealing with multi-label classification scenarios. The scope of this study decided to implement and investigate the PubMedBERT model which the Mekkes study found to be the best-performing model in terms of micro-F1 score.

Furthermore, due to limited computing resources (after several attempts ending in crashes or malfunctioning), the training dataset was reduced from the original 8000+ samples to 2000 (with a randomized 80:20% split for training and validation). The data was first preprocessed to remove any indices with missing values before reupdating the dataset to train and test in both BoW and BERT NLP architectures. I thought it would be interesting to explore the parameters of BiomedBERT to the extent they were available, so the following architectural parameters were employed in the simulated NLP model:

Table 2. BiomedBERT model configuration (as per the HuggingFace repository, Gu et al. 2020)

Parameter	**Description**
Attention probability, Dropout probability	0.1
Hidden activation function	Gaussian error linear unit (GELU)
Hidden dropout probability	0.1
Hidden size (number of neurons)	786
Initializer range	0.02
Intermediate size	3072
Max position embeddings	512
Number of attention heads	12
Number of hidden layers	12
Type of vocab size	2
Vocabulary size	30522

The Adam optimizer used a learning rate of 5e-5.

Outline of Investigation

The pretraining stages of this model consisted of masking random components of training sentences to achieve reconstruction of the original sentence. Two models were compared; bag-of-words as a stand-in for classical NLP models and the transformer-based model which employs a neural network to update its internal parameters during training.

The original model also used 50 epochs to train (or full passes through the dataset); this was beyond the computational capacity of this study. Employment of 5 epochs of training will be attempted with considerations of optimizing training batch sizes – based on computational outcomes this may be subject to change. Both models used an 80:20 split of randomized samples for training and validation (as would be consistent with the Mekkes model). The stratification method will be excluded to try and remove the aspect of manipulating training which may not directly correlate to how brains process information.

Models were run using Google Colab in the Python programming language (v. 3.11). HuggingFace models were employed (imported via the Transformers library, v. 4.40.0), along with the necessary packages from Pandas (2.0.3), Numpy (1.25.2), PyTorch (v. 2.2.1+cu121), Matplotlib (v. 3.7.1) and Scikit-learn (1.2.2) dependencies. Due to the set-up complexity of the original models and the simplified objectives explored in this study, portions of the implemented models were generated and troubleshooted in part by ChatGPT (OpenAI 2024).

Results

An initial attempt at running the full corpus of NLP training data (8227 sentences) cost a total compute time of about 8 hours when running the Biomed-BERT model – unfortunately this crashed at 5th training epoch so a complete output could not be computed for this trial. For this reason, the training data was broken into a subset for feasible simulation (2000 cases to test the model out).

For example, for the initial total split dataset of 8277 (the post-processing size):

- Training size = 6621 samples will require roughly 1000 iterations to complete all 5 epochs at a batch size of 33
- Testing size = 1656 will need 50 iterations

Compute time was the first differentiable aspect between the two models; it drastically increased between the BoW (roughly 4 minutes) and transformer-based model, the latter of which depended on the size of the input dataset.

Upon implementing the BoW and BioBERT models, both output a classification table consisting of 90 attributes with a classification report detailing precision, recall and f1-scores for micro, macro, weighted and sample averages. The relevant criteria are displayed below:

Table 3.0. Micro-metrics of tested BoW and BERT models

NLP method // Metric	BoW	$BERT_{1000\ samples\ \&\ 3\ training\ epochs}$	$BERT_{4000\ samples\ \&\ 1\ training\ epoch}$
Precision	0.97	0.97	0.73
Recall	0.48	0.44	0.58
f1-score	0.64	0.60	0.65

Table 3.1. Sample-metrics of tested BoW and BERT models

NLP method // Metric	BoW	$BERT_{1000\ samples\ \&\ 3\ training\ epochs}$	$BERT_{4000\ samples\ \&\ 1\ training\ epoch}$
Precision	0.62	0.57	0.73
Recall	0.53	0.48	0.60
f1-score	0.56	0.51	0.64

The micro-precision, recall and f1-score are reported in Table 3.0 as a comparison to the metric of choice in the Mekkes model. Sample metrics are also reported (Table 3.1) as it is another commonly-used metric where computed scores are averaged across each instance, and seem like a relevant consideration in this case as inputs can belong to multiple classes.

The most striking effect of comparing micro to sample metrics is the fall in precision marks; this may be due to an under sampling by virtue of the randomized sample pools. During BiomedBERT training, it was observed that NLP_{1000} got down to a loss of 0.12, while NLP_{4000} (with 1 epoch) went even further to a loss of 0.07 (batch size was kept constant). No unstable gradients were observed at this size either. Interestingly, an increased training set showed enhanced training performance even though the testing outcomes were comparable across all three models.

Another interesting result was the need to optimize batch sizing when it came to employing the BioBERT model. As a larger model there was an increase of computing resources and time spent on training and testing; a straightforward workaround to conserve memory

resources was to introduce data in batches. This needed careful consideration as well, due to small batches leading to more frequent parameter updating and introduce unstable gradients in the minimization of loss steps, as well as slowing training down. This was observed when testing with a batch size of 16 (roughly 400 iterations to complete an epoch) – within the first epoch (of 5) the loss function jumped to 2.35 after stabilizing at 0.4 in the middle of the epoch, eventually slowly decaying back to 0.4. In initial trials (with a smaller randomized sample), an optimal batch size of 33 samples was determined (which was found to be the optimal level in terms of efficiency while maintaining available RAM).

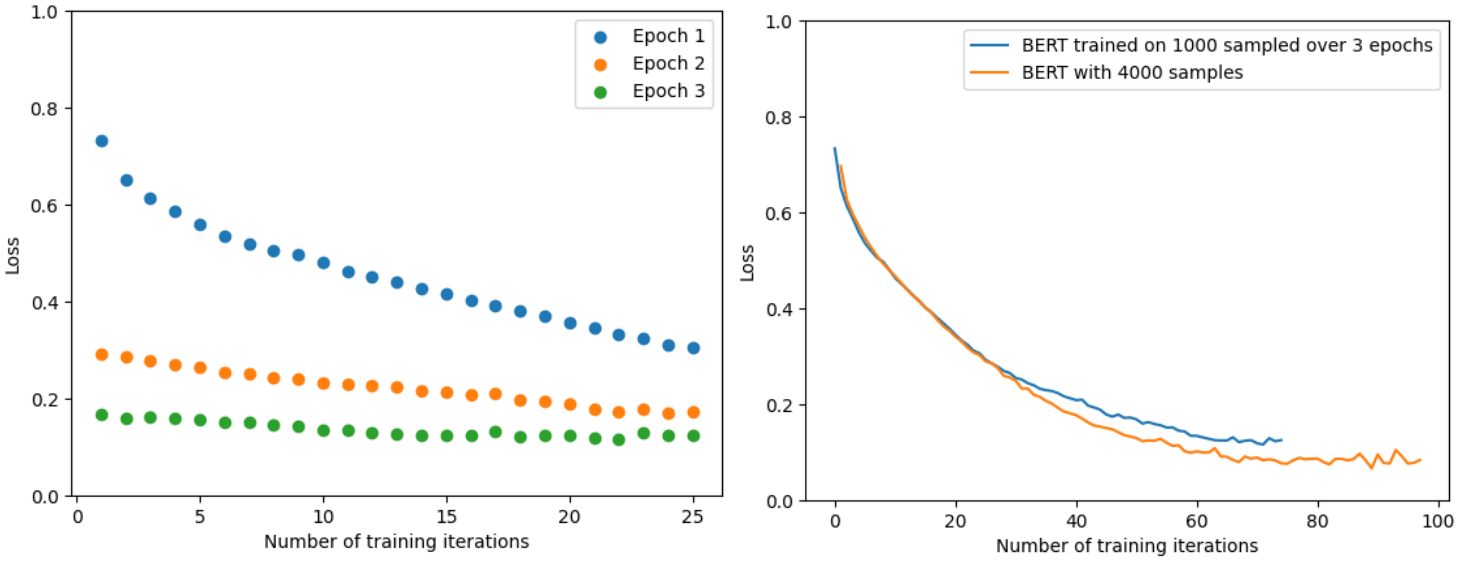

Figure 1.0. BERT loss convergences for a) across epochs when trained on 1000 samples and b) between 1000 and 4000 samples over 3 and 1 epochs, respectively.

In the validation portion of the model, certain trials showed that the BoW model unexpectedly performed better than the BERT model on a drastic level (f1-scores and precision scores were computed to be 0.0 for BERT), indicating that something may not have been implemented correctly during the testing portion of the model. This caveat is mentioned to indicate a certain level of scrutiny for these results so a next step would involve reproducing the data observed.

Discussion

Overview of Findings & Limitations

It is crucial that we understand how doctors and physicians understand clinical histories and trajectories as attributes of neuropathological outcomes and disorders. This study was interested in exploring several aspects of NLP architectures, namely transformer and non-transformer-based methods as well as the nature of training of transformer-based methods within the bounds of trade-offs between batch sizes, epochs and sample sizes. Due to computational limits (and programming skill levels) high-level versions of the published models were explored to determine how these methods behave in different training environments and with different training sets.

By removing the stratification layer of preprocessing employed by Mekkes et al. I was curious to see the outcome when distributed sampling is not introduced in training of NLPs. Essentially, to what extent can different NLP models maintain their performance when processing subpar or nonrepresentative datasets? As expected, performance metrics decreased to roughly 60% relative to the published model's range of 80-93% (Mekkes et al. 2024).

Inputs to the brain are not naturally stratified; we don't choose how we're exposed to groups of data and can't consciously forget something we've learned to reuse that data to be tested on. The use of this sampling technique was interesting to me because it doesn't seem to correlate naturally to how the brain likely processes information, yet its inclusion greatly benefits NLPs. Of course, on one hand, without employing the entirety of the dataset, it is difficult to conclude the validity of these results, but on the other, the incompleteness correlates to a valid limitation about the nature of medical data.

Mekkes et al. believe that medical summaries are innately incomplete due to the intrinsic nature of medical records only encompassing the primary focus of each visit. They attribute this as a sampling problem which is often characteristic to biomedical applications of NLP models (Feczko et al. 2019). This study used a method of temporally tracking clinical history to build disease manifestation records on the basis of this issue (Xie et al. 2022). Based on this, what may be interesting to pursue further is comparing diagnostic outcomes from physicians exploring medical data from the approach of building clinical trajectories and how well the NLP outputs align with the medical diagnoses, since the current state of development is compares the hypothetical situation of having labelled medical data to reality where data can be incomplete and stochastic.

Additionally, this paper was supervised learning method while other AI methods use unsupervised learning methods such as clustering, dimensionality reduction, association learning rules such as in KRISSBERT (knowledge-rich self-supervision bidirectional encoder

representations transformers). It would have been nice to be able to alter different activation functions within the BERT model instead but those were unmodifiable due to their predefinition from the pretrained model. BiomedBERT is known to use the GELU nonlinear activation function:

$$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}(x/\sqrt{2})\right],$$

where erf is the error function

As depicted in the following graph, GELU behaves as a smoother ReLU (rectified linear unit) due to its incorporation of the Gaussian normal distribution function.

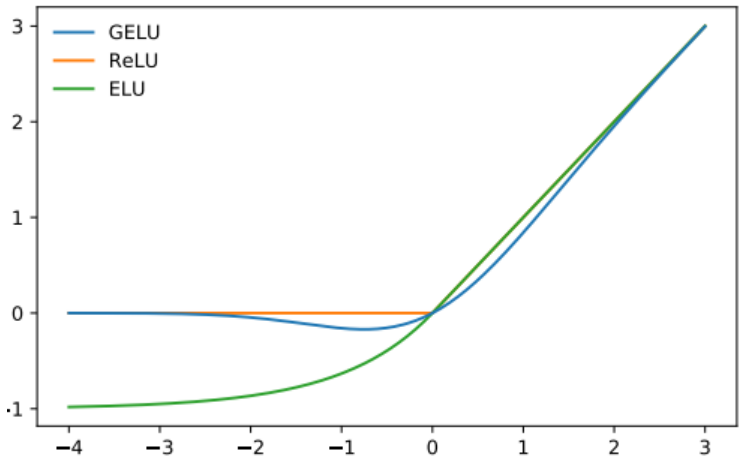

Figure 2.0. Comparisons of linear unit function derivatives (Zanella & Toussaint 2022)

In the current model implementation, only downstream activation functions for the output computation could be altered as the internal parameters of the BiomedBERT model were unmodifiable, so that would be a first step on focusing efforts of tuning activation functions. It also leads to questioning operating under the assumption of a potential role and purpose of having multiple or different training and testing activation functions for NLPs.

One aspect that would have been nice to implement is usage of the penalized hyperbolic tanh function (ptanh) which has been reported to perform well in NLP applications when compared to other activation functions like ReLU (Kunc & Kléma 2024).

$$f(z) = \begin{cases} \tanh(z), & z \geq 0, \\ \dfrac{\tanh(z)}{a}, & z < 0, \end{cases}$$

I was unable to identify this in the suite of activation functions available in the PyTorch library, though this could be introduced through a penalty term into a customized tanh activation function.

Analysis of evaluation metrics

The scikit-learn classification report produces a summary of precision, recall and f1-scores, as well as an overview on support (which is the number of samples within each class). Metrics for multi-label classifiers are an interesting area highlighting the importance of considering the statistics of your dataset; as Mekkes et al. believed their dataset to be imbalanced, they assessed performance using micro-F1 (which is the harmonic mean of precision and recall).

With respect to metrics for multi-classifiers, though f1-scores are widely-reported, a higher score generally does but may not always represent a better-performing model (as is the primary evaluator reported by Mekkes et al.). As it is a harmonic mean between precision and recall, there is a trade-off between tuning for greater precision which can often reduce the recalling ability.

On the basis of the confusion matrix, the correctness of outputs is classified as true or false positives, and true or false negatives based on whether the observed outcome matches up to what was predicted (Zanella & Toussaint 2022). The ratio of these outcomes can be described as the evaluation metrics used to rate the performance of ML models, namely the precision, recall and f1-score of a model:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$f1 - score = \frac{2 * precision * recall}{precision + recall}$$

Furthermore, classification tasks for where inputs belong to exactly a single class (either binary or multi-class) see f1-scores equivalent to their accuracy as in these scenarios they are independent from true negatives - this is not the case for multi-label situations.

NLPs and Biological Plausibility

When considering the actual translation of NLP models and their mimicry of the human brain, we delve into the biological plausibility of machine learning's ability to process language.

When using vector-based methods such as bag-of-words, we lose the context dependency between words but we regain that when we pivot to transformer-based methods like BERT, which do encompass representations between words in a sentence (Alsentzer et al. 2019).

Similarly, the brain uses neural networks to integrate language and meaning, primarily in the posterior superior temporal lobe where Wernicke's area is housed. This brain region is associated with language comprehension (both written and spoken), which best emulates the task of the multilabel text classification models studied in this paper. Wernicke's area (also called Brodmann Area 22) is responsible for retrieving word meanings and making semantic associations, which transformer-based models also aim to capture in their fine-tuning mechanisms (Schrimpf et al. 2021).

Some additional brain region parallels include the inferior parietal lobule which performs some feature extraction and visual mapping of written words to their meanings in for semantic integration, which also happens to align with text-trained NLPs (Caucheteux & King 2022). Caucheteux & King observe that brains and NLPs do converge partially based on the mapping of model activation functions onto brain responses to visual text, identifying that the middle-hidden layers (which are inaccessible to us in the BiomedBERT model) estimate this representation best.

Another similar study also found a striking resemblance between transformer-based NLP methods and the hierarchical language processing in the brain; when they compared electrocorticography measurements of language areas in the brain to GPT2-XL, there were corresponding activity patterns between the sequence of embeddings across GPT2-XL's layers and the sequence of neural activity in the associative temporal brain regions (Goldstein et al. 2023). An interesting distinction, however, was the observation of brains to process unexpected (contextless, or "surprising") words that couldn't have been predicted; this property was not carried over to the NLP embedding sequences.

The neat thing is the similarity in findings that these studies are amassing between one another; Caucheteux & King report consistency in their findings with Schrimpf et al. when it comes to how masking in deep learning algorithms can be mapped to how the brain represents language in a hierarchical fashion allowing predictions of masked and contextual text. One key difference is the lack of a direct parallel between the brain's ability to facilitate communication between the Wernicke's and Broca's areas (which integrate auditory information for the generation of coherent output), and the way NLP methods generate outputs as all they can use are the input, output, and hidden layers.

Additionally, the brain integrates several distributed networks over various regions for linguistic processing, ranging from phonetics to syntax to grammar in addition to the overall meaning of the text to conclude a message. Both NLP models and the brain generalize tasks for eventual fine-tuning to novel situations since adaptation is a natural part of living systems

(Caucheteux & King 2022). The brain's ability to generalize when it comes to language supersedes that of NLPs. The brain also trumps NLP capabilities of speed and efficiency when it comes to processing volumes of input – upon learning how to read this can become an unconscious effort of processing input (which would require an involved pretraining effort on the NLP's part).

Real human brains do seem more akin to transfer based learning methods; we are constantly exposed to information that we process, interpret and inform ourselves to bring forth in our future assessments and task completion (both consciously and subconsciously).

Acknowledgements

Independent of this final project, I wanted to thank the three of you (Michael, Terry and Trevor) for your unwavering support throughout the term. I genuinely enjoyed having the opportunity to learn from the three of you and appreciated the efforts you put in to make this possible for my academic career. I've never taken a class with professors and teaching staff like this before, and it was a real privilege to get to experience that.

Though I originally didn't even expect to get a functional model working for this assignment, your accommodation waiving that requirement alone gave me the peace of mind to organize myself to be able to achieve some semblance of a model after all which I was able to explore through this work. Thank you for providing me with this space for learning something completely out of my comfort zone – thanks to you I can say that I'm very much looking forward to continuing on this path in the future.

All the very best,

Aaisha

References

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly Available Clinical BERT Embeddings (arXiv:1904.03323). arXiv. https://doi.org/10.48550/arXiv.1904.03323

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. Communications Biology, 5(1), 1–10. https://doi.org/10.1038/s42003-022-03036-1

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., & Fair, D. A. (2019). The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes. Trends in cognitive sciences, 23(7), 584–601. https://doi.org/10.1016/j.tics.2019.03.009

Goldstein, A., Ham, E., Nastase, S. A., Zada, Z., Grinstein-Dabus, A., Aubrey, B., Schain, M., Gazula, H., Feder, A., Doyle, W., Devore, S., Dugan, P., Friedman, D., Brenner, M., Hassidim, A., Devinsky, O., Flinker, A., Levy, O., & Hasson, U. (2023). Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain (p. 2022.07.11.499562). bioRxiv. https://doi.org/10.1101/2022.07.11.499562

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Naumann, T., Gao, J., & Poon, H. (2020). PubmedBERT. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. http://arXiv:2007.15779

Kunc, V., & Kléma, J. (2024). Three Decades of Activations: A Comprehensive Survey of 400 Activation Functions for Neural Networks (arXiv:2402.09092). arXiv. http://arxiv.org/abs/2402.09092

Lehman, E., Hernandez, E., Mahajan, D., Wulff, J., Smith, M. J., Ziegler, Z., Nadler, D., Szolovits, P., Johnson, A., & Alsentzer, E. (2023). Do We Still Need Clinical Language Models? (arXiv:2302.08091). arXiv. http://arxiv.org/abs/2302.08091

OpenAI. (2024). ChatGPT (April 2024 version) [Large language model]. https://chat.openai.com/chat

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2018). Scikit-learn: Machine Learning in Python (arXiv:1201.0490). arXiv. https://doi.org/10.48550/arXiv.1201.0490

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

Sechidis, K., Tsoumakas, G., & Vlahavas, I.P. (2011). On the Stratification of Multi-label Data. ECML/PKDD. https://link.springer.com/chapter/10.1007/978-3-642-23808-6_10

trent-b/iterative-stratification: scikit-learn cross validators for iterative stratification of multilabel data. (2022). GitHub. https://github.com/trent-b/iterative-stratification

Xie, F., Yuan, H., Ning, Y., Ong, M. E. H., Feng, M., Hsu, W., Chakraborty, B., & Liu, N. (2022). Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. Journal of biomedical informatics, 126, 103980. https://doi.org/10.1016/j.jbi.2021.103980

Zanella, L., & Toussaint, Y. (2022). Fine-tuning Pre-trained Transformer Language Models for Biomedical Event Trigger Detection. Atelier DL4NLP - Extraction et Gestion de Connaissances (EGC). https://hal.science/hal-03984783

Appendix

Code availability: the Google Colab file can be made available if interested (it is currently in a state of reformatting for clarity)

Supplementary data:

Table 4. Groupings of attributes for computational classification

Attribute grouping	Attributes
Balance problems overall	Ataxia, Balance problems, Frequent falls, Loss of coordination, Nystagmus, Vertigo
Behavioral changes	Aggressive behavior, Agitation, Changed behavior/personality, Changed moods or emotions, Concentration problems, Lack of insight, Loss of sympathy / empathy
Cognitive slowness	Apathy / inertia, Bradyphrenia, Fatigue, Lack of initiative
Communication impairment	Aphasia, Communication problems, Dysarthria, Impaired comprehension, Language impairment, Wordfinding problems
Dementia overall	Cognitive decline, Dementia
(Dis)inhibition overall	Disinhibition, Loss of decorum
Disorientation overall	Confusion, Disorientation, Wandering
Executive dysfunction overall	Apraxia, Executive function disorders, Lack of planning / organization / overview
Memory impairment overall	Amnesia, Confabulations, Façade behavior, Forgetfulness, Headturning sign, Impaired recognition, Imprinting disturbances, Memory impairment
Mobility problems	Mobility problems, Unspecified disturbed gait patterns
Motor deficit	Decreased (fine) motor skills, Fasciculations, Hyperreflexia and other pathological reflexes, Muscular weakness, Spasticity
Neurological symptom not classified	Frontal release signs, Headache / migraine, Orthostatic hypotension, Seizures
Parkinsonism overall	Bradykinesia, Facial masking, Parkinsonism, Rigidity, Tremor
Physical Decline	Admission to nursing home, Day care, Declined / deteriorated health, Help in ADL
Psychiatric symptoms	Anxiety, Compulsive behavior, Delirium, Delusions, Depressed mood, Hallucinations, Hyperorality, Mania, Paranoia and suspiciousness, Psychiatric admissions, Psychosis, Stress, Suicidal ideation
Reduction of food intake	Cachexia, Reduced oral intake, Swallowing problems / dysphagia, Weight loss

Sensory deficits	Hearing problems, Negative sensory symptoms, Olfactory and gustatory dysfunction, Positive sensory symptoms, Visual problems
Sleep & wake cycle abnormalities	Day/night rhythm disturbances, Restlessness, Sleep disturbances, Vivid dreaming
Urinary and bowel dysfunction	Constipation, Urinary incontinence, Urinary problems (other)

Table 5. Abbreviations of progressive neurodegenerative diseases focused in Mekke et al. study for mapping clinical trajectories

Abbreviation	**Disorder**
AD	Alzheimer's disease
PD/PDD	Parkinson's disease/Parkinson's disease with dementia
VD	Vascular dementia
FTD	Frontotemporal dementia
DLB	Dementia with Lewy bodies
AD-DLB	Alzheimer's disease and dementia with Lewy bodies
ATAXIA	Cerebellar ataxia
MND	Motor neuron disease
PSP	Progressive supranuclear palsy
MS	Multiple sclerosis
MSA	Multiple system atrophy